(12) UK Patent Application (19) GB (11) 2 386 282 (13) A

(43) Date of A Publication 10.09.2003

(21) Application No 0205105.0

(22) Date of Filing 05.03.2002

(71) Applicant(s)
PA Consulting Services Ltd
(Incorporated in the United Kingdom)
Cambridge Technology Centre,
MELBOURN, Hertfordshire, SG8 6DP,
United Kingdom

(72) Inventor(s)
Simon Charles Durrant
John David Ainsworth

(74) Agent and/or Address for Service
Withers & Rogers
Goldings House, 2 Hays Lane, LONDON,
SE1 2HW, United Kingdom

(51) INT CL[7]
H04Q 11/04

(52) UK CL (Edition V)
H4K KOT KTKX
H4L LRRMW L205 L213

(56) Documents Cited
EP 0734195 A2 **WO 2001/089234 A2**
US 5892754 A

(58) Field of Search
UK CL (Edition T) **H4K** KOD8 KOT KTKA KTKX, **H4L** LDGP LRRMS LRRMW
INT CL[7] **H04L** 12/56 29/06, **H04Q** 7/22 11/04
Other: **On-Line - EPODOC, JAPIO, WPI**

(54) Abstract Title
Allocating shared resources in a packet data communications network

(57) The allocation of shared transmission resources are optimised between two or more packet data streams of applications executing in a device in a packet data communications network. The device includes a transmission resource manager function. Each application specifies a range of acceptable values for one or more of the parameters that determine the transmission quality of its data stream, and the transmission resource manager function allocates the available transmission resources to all the data streams in dependence on the range of acceptable values supplied by each application. Preferably, the application supplies for each parameter a range in the form of a target value, which is the preferred value for the application, and a limit value, which is the minimum acceptable value to the application. Transmission resources may be allocated in such a way as to maximise the number of active data streams. May be applied to GPRS and UMTS radio communication systems.

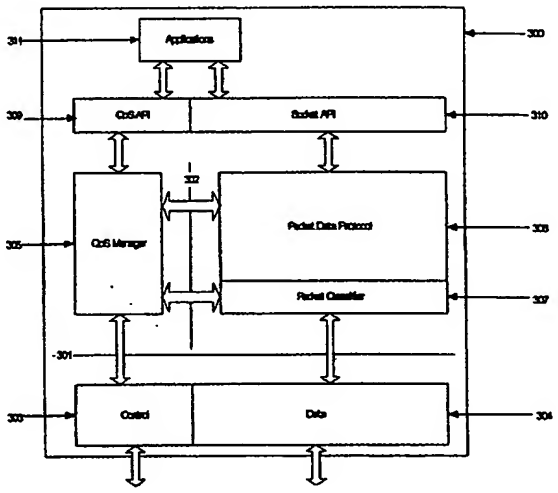

Figure 3

GB 2 386 282 A

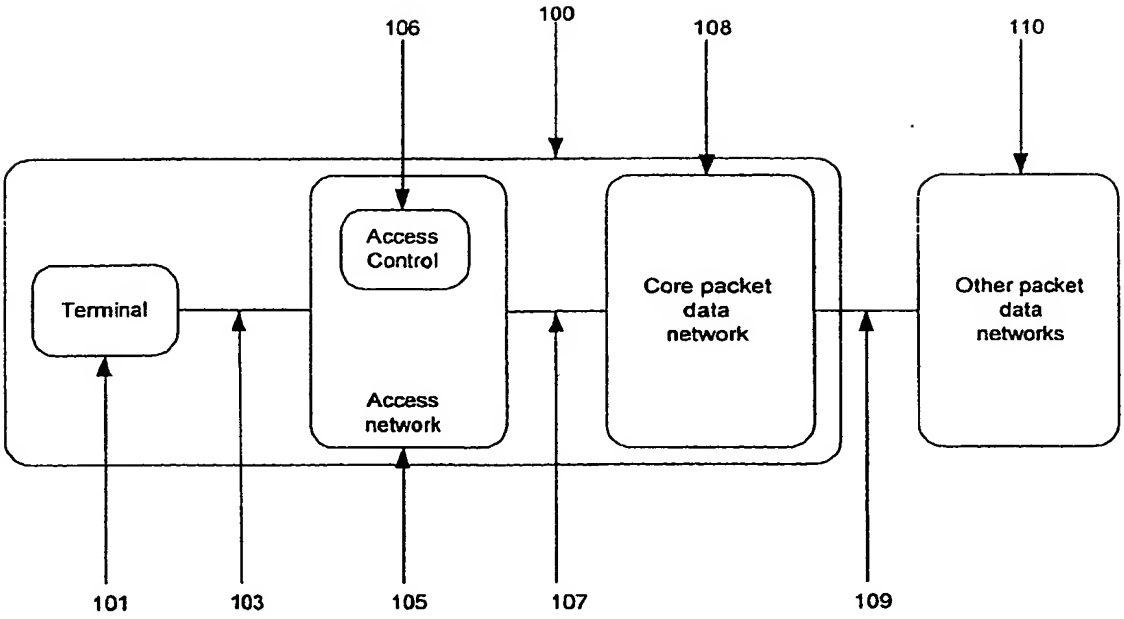
106
100
108
110
Access Control
Terminal
Core packet data network
Other packet data networks
Access network
101
103
105
107
109

Figure 1

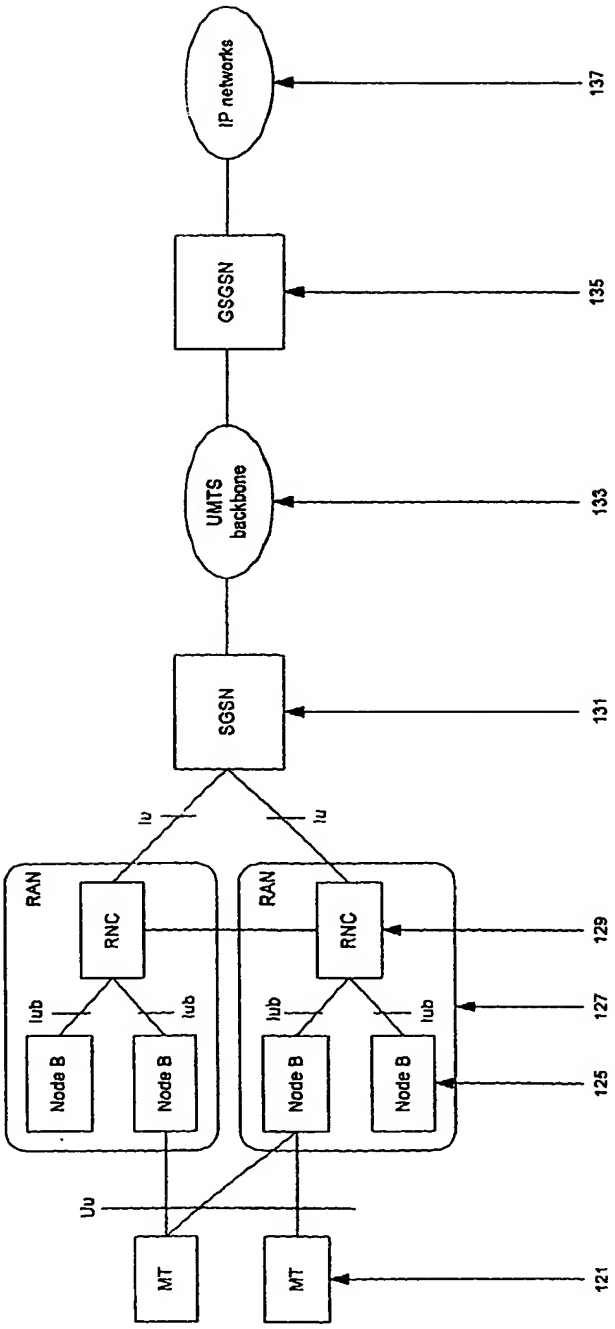
IP networks
137
GSGSN
135
UMTS backbone
133
SGSN
131
Iu
Iu
RAN
RNC
129
Iub
Iub
Node B
Node B
127
125
RAN
RNC
Iub
Iub
Node B
Node B
Uu
MT
MT
121

Figure 2

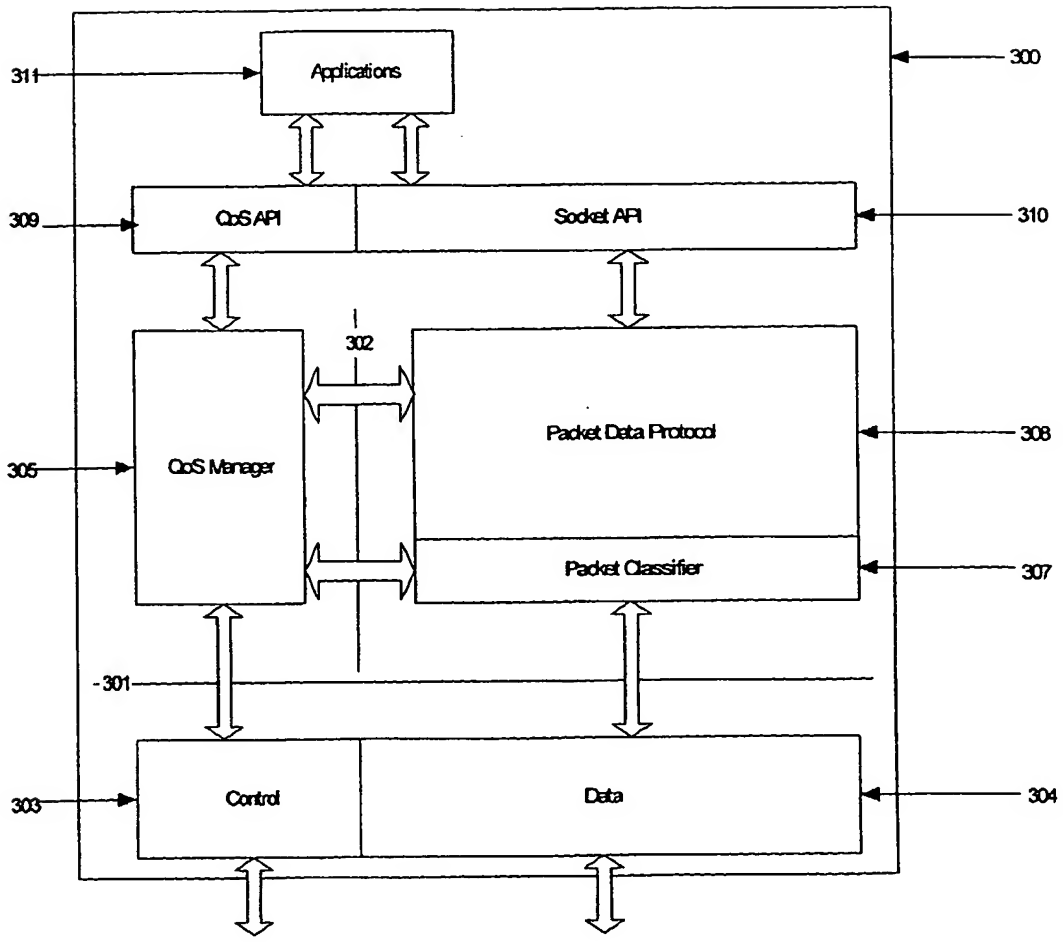
Applications
311
300
QoS API
309
Socket API
310
302
Packet Data Protocol
308
QoS Manager
305
Packet Classifier
307
301
Control
303
Data
304

Figure 3

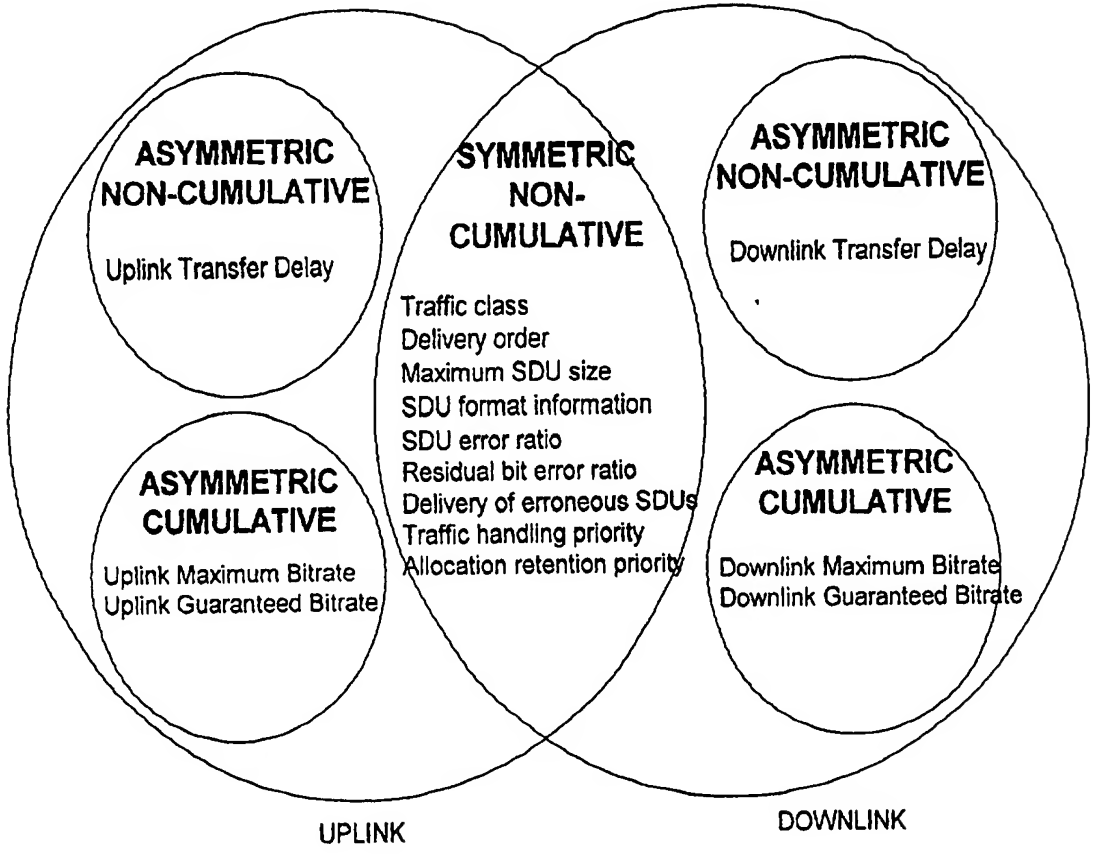
ASYMMETRIC
NON-CUMULATIVE
Uplink Transfer Delay
ASYMMETRIC
CUMULATIVE
Uplink Maximum Bitrate
Uplink Guaranteed Bitrate
SYMMETRIC
NON-
CUMULATIVE
Traffic class
Delivery order
Maximum SDU size
SDU format information
SDU error ratio
Residual bit error ratio
Delivery of erroneous SDUs
Traffic handling priority
Allocation retention priority
ASYMMETRIC
NON-CUMULATIVE
Downlink Transfer Delay
ASYMMETRIC
CUMULATIVE
Downlink Maximum Bitrate
Downlink Guaranteed Bitrate
UPLINK
DOWNLINK

Figure 4

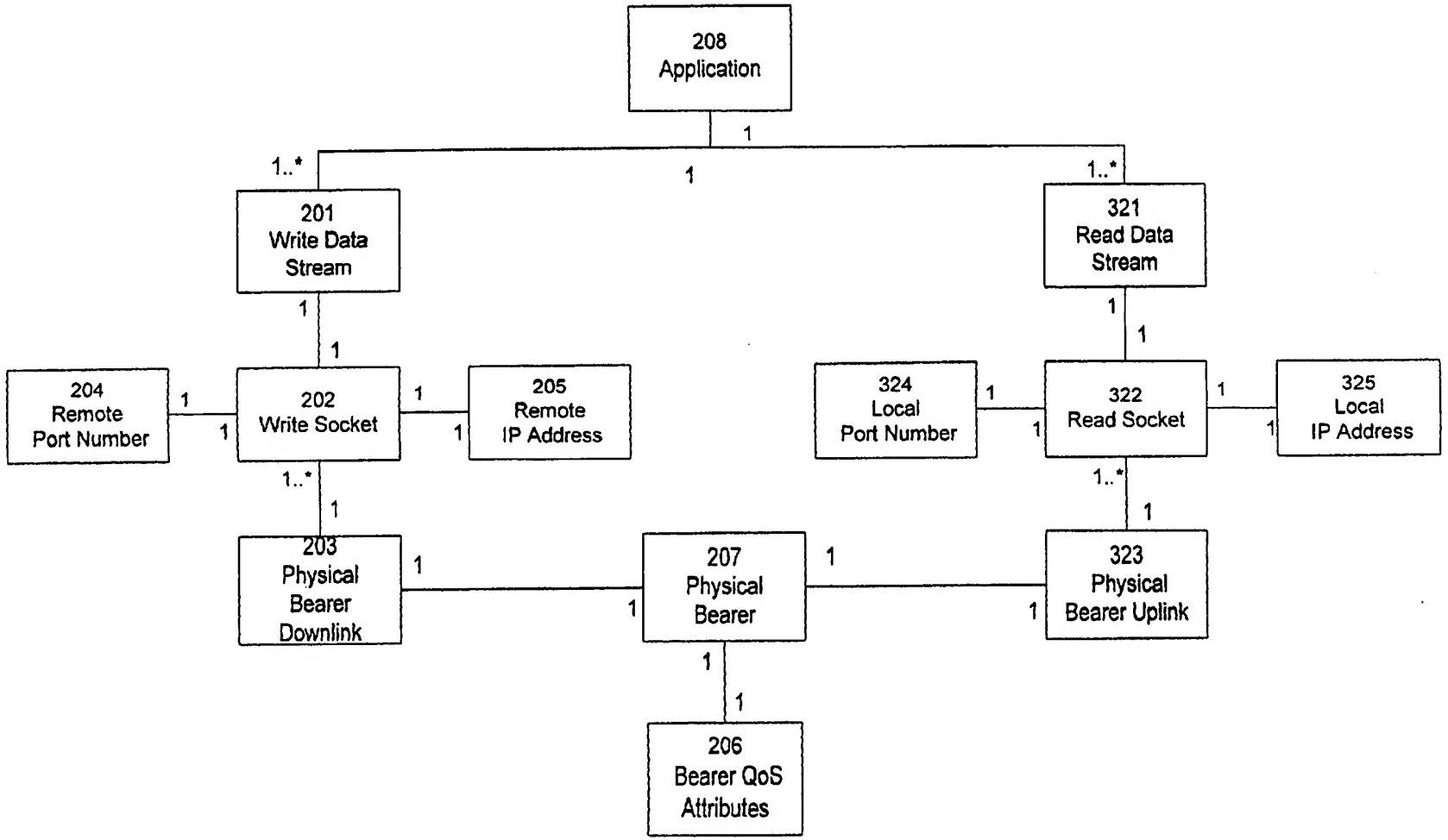
208
Application
1
1..*
1
1..*
201
Write Data
Stream
321
Read Data
Stream
1
1
1
1
204
Remote
Port Number
1
1
202
Write Socket
1
1
205
Remote
IP Address
324
Local
Port Number
1
1
322
Read Socket
1
1
325
Local
IP Address
1..*
1
1..*
1
203
Physical
Bearer
Downlink
1
1
207
Physical
Bearer
1
1
323
Physical
Bearer Uplink
1
1
206
Bearer QoS
Attributes

Figure 5

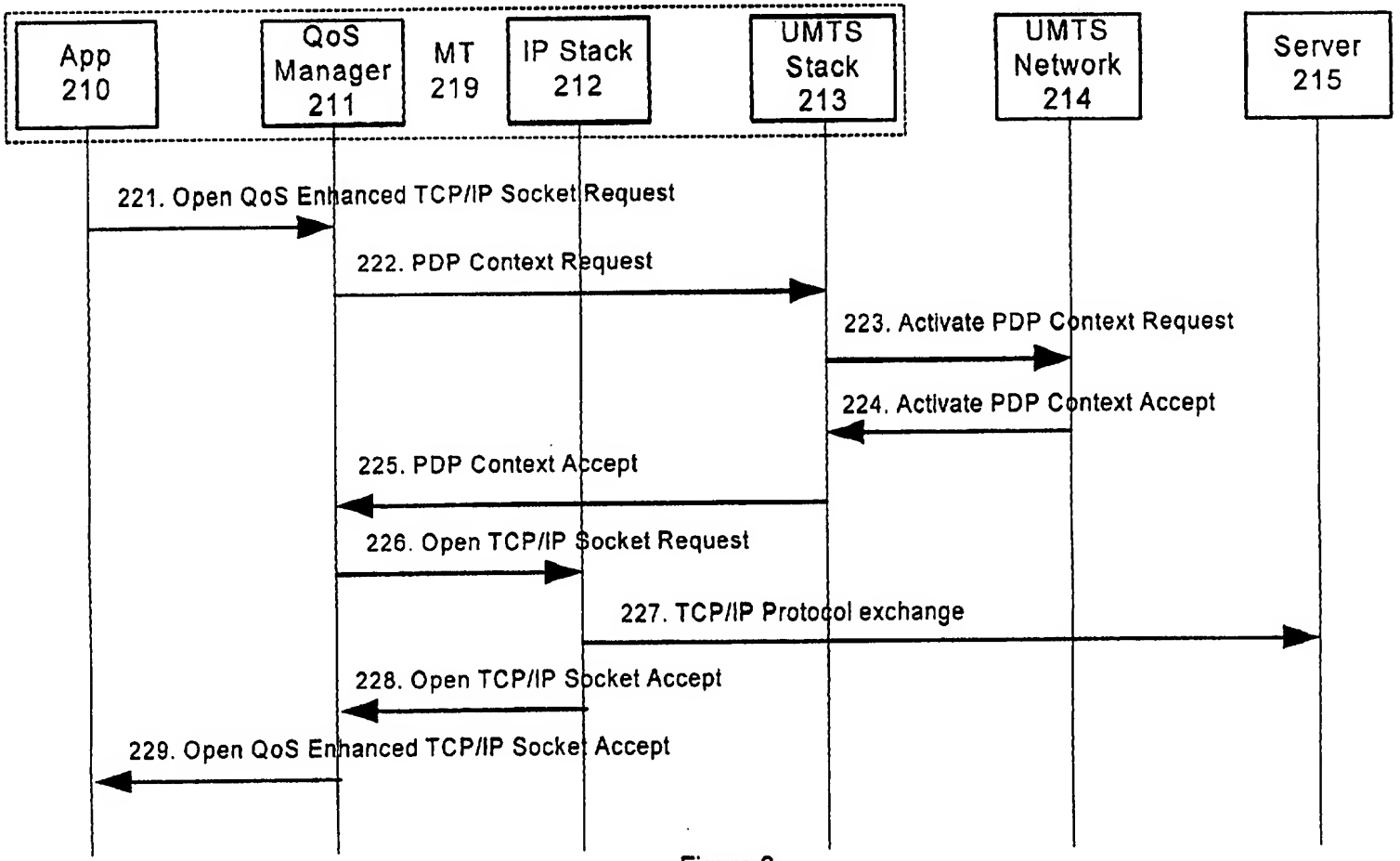
App 210
QoS Manager 211
MT 219
IP Stack 212
UMTS Stack 213
UMTS Network 214
Server 215
221. Open QoS Enhanced TCP/IP Socket Request
222. PDP Context Request
223. Activate PDP Context Request
224. Activate PDP Context Accept
225. PDP Context Accept
226. Open TCP/IP Socket Request
227. TCP/IP Protocol exchange
228. Open TCP/IP Socket Accept
229. Open QoS Enhanced TCP/IP Socket Accept

Figure 6

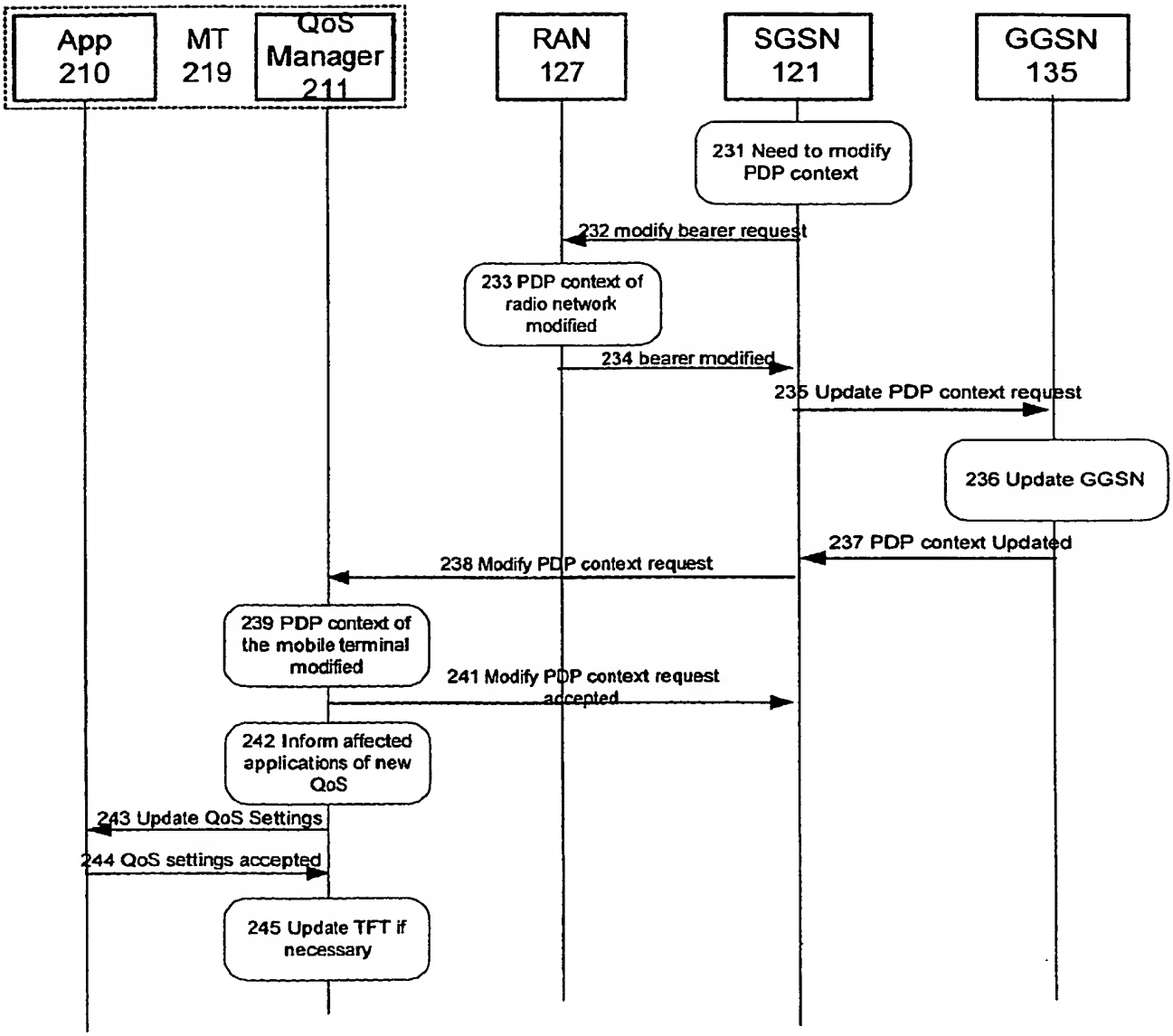
App 210
MT 219
QoS Manager 211
RAN 127
SGSN 121
GGSN 135
231 Need to modify PDP context
232 modify bearer request
233 PDP context of radio network modified
234 bearer modified
235 Update PDP context request
236 Update GGSN
237 PDP context Updated
238 Modify PDP context request
239 PDP context of the mobile terminal modified
241 Modify PDP context request accepted
242 Inform affected applications of new QoS
243 Update QoS Settings
244 QoS settings accepted
245 Update TFT if necessary

Figure 7

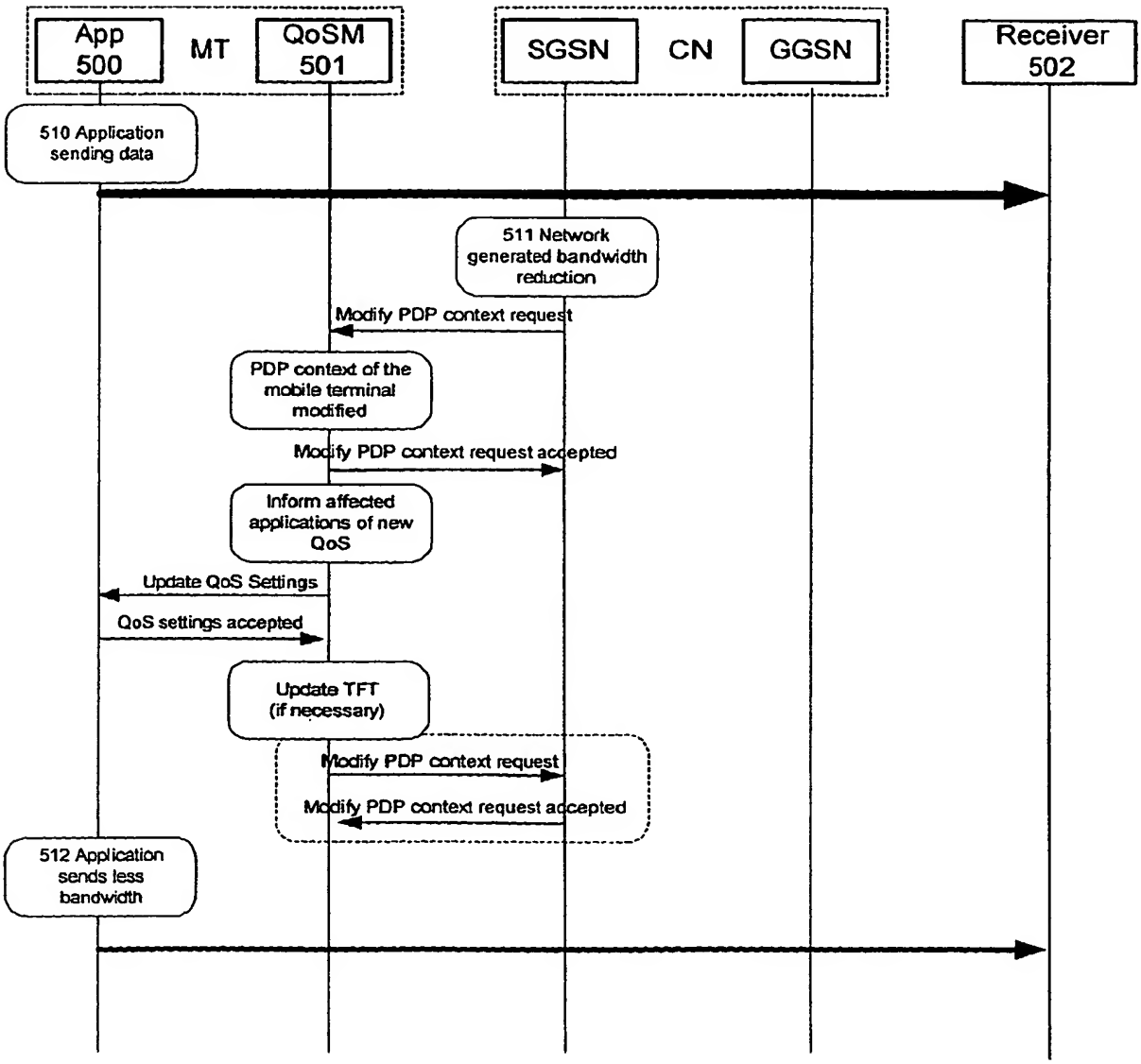
App 500
MT
QoSM 501
SGSN
CN
GGSN
Receiver 502
510 Application sending data
511 Network generated bandwidth reduction
Modify PDP context request
PDP context of the mobile terminal modified
Modify PDP context request accepted
Inform affected applications of new QoS
Update QoS Settings
QoS settings accepted
Update TFT (if necessary)
Modify PDP context request
Modify PDP context request accepted
512 Application sends less bandwidth

Figure 8

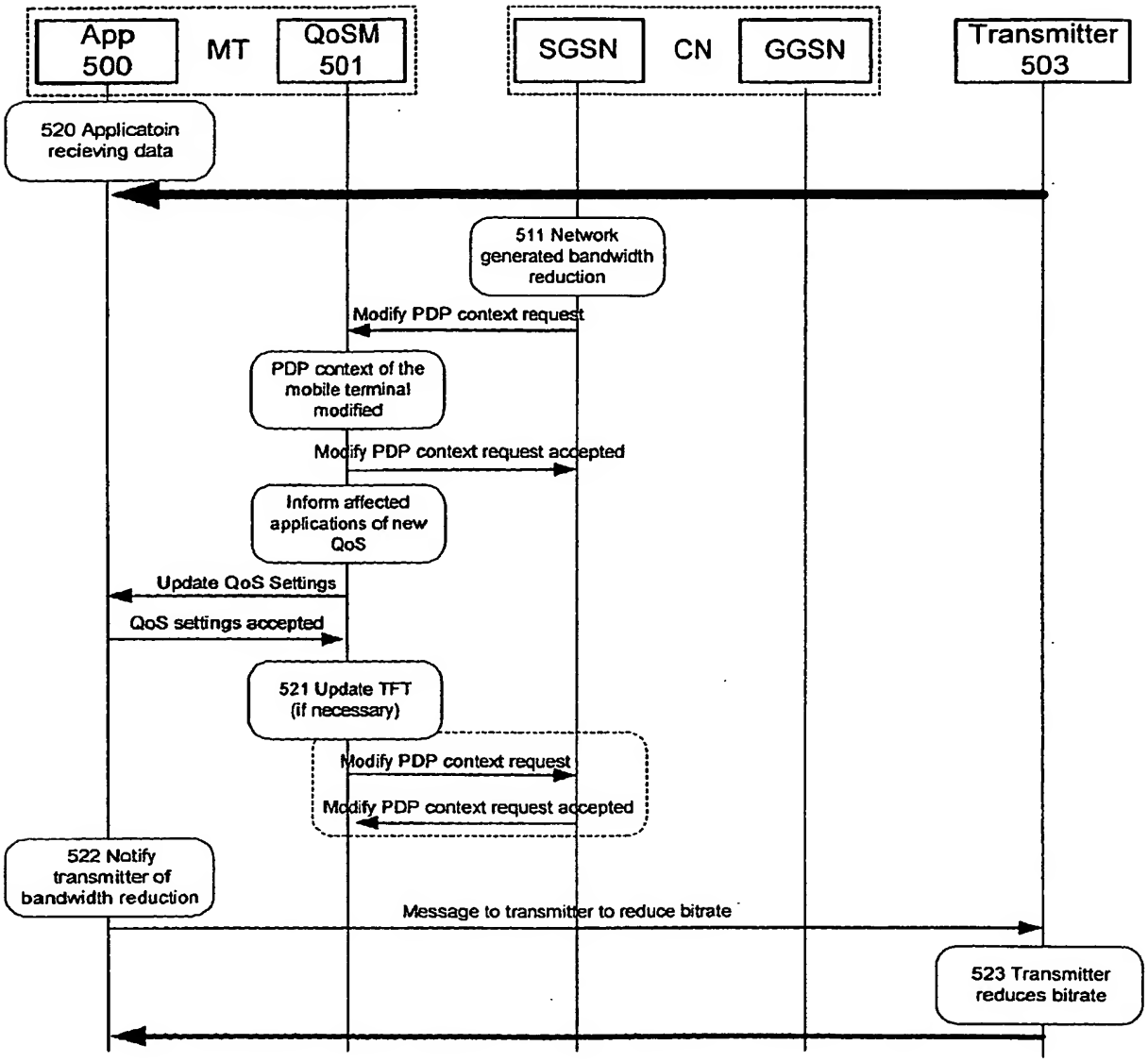
App 500
MT
QoSM 501
SGSN
CN
GGSN
Transmitter 503
520 Applicatoin recieving data
511 Network generated bandwidth reduction
Modify PDP context request
PDP context of the mobile terminal modified
Modify PDP context request accepted
Inform affected applications of new QoS
Update QoS Settings
QoS settings accepted
521 Update TFT (if necessary)
Modify PDP context request
Modify PDP context request accepted
522 Notify transmitter of bandwidth reduction
Message to transmitter to reduce bitrate
523 Transmitter reduces bitrate

Figure 9

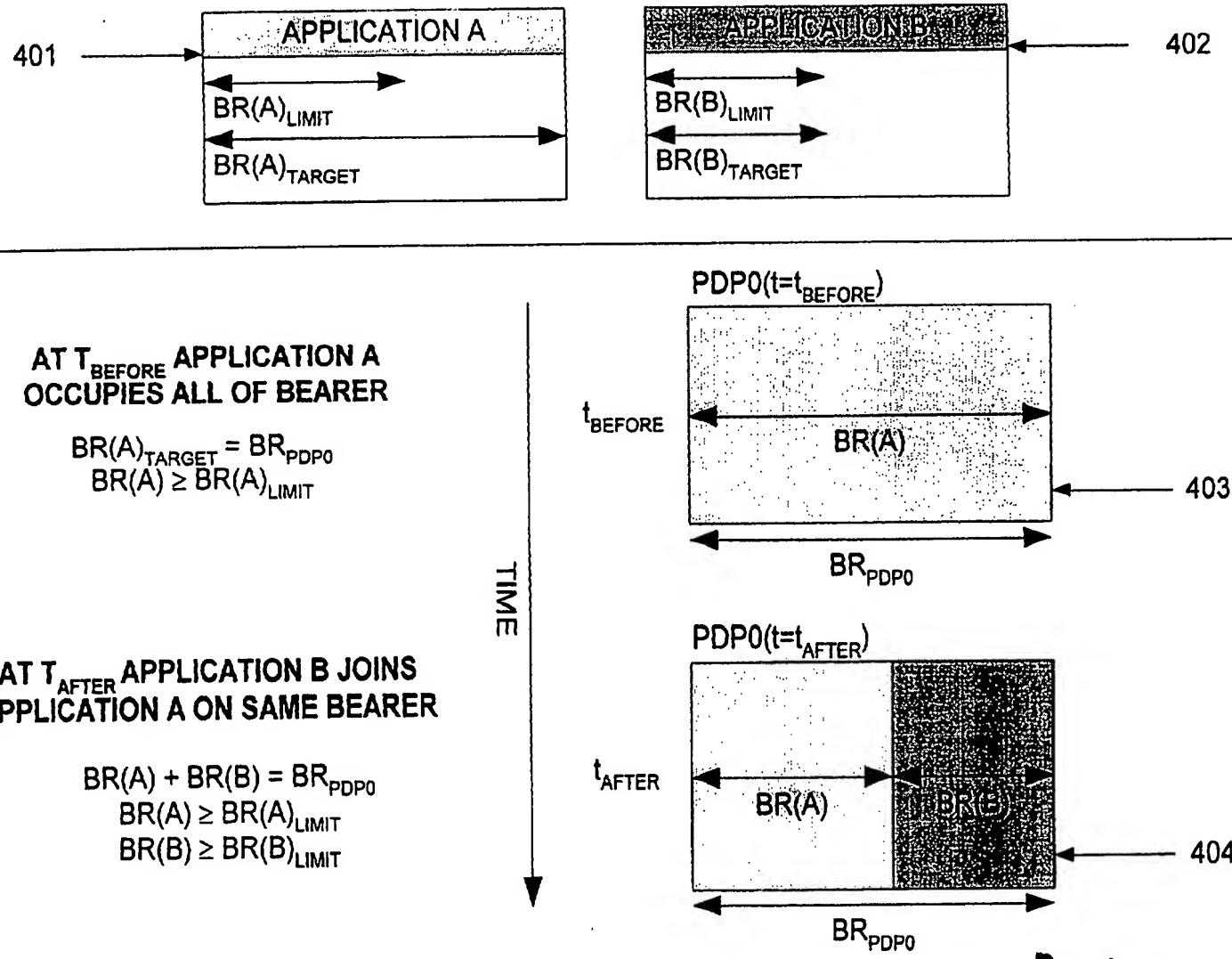

Figure 10

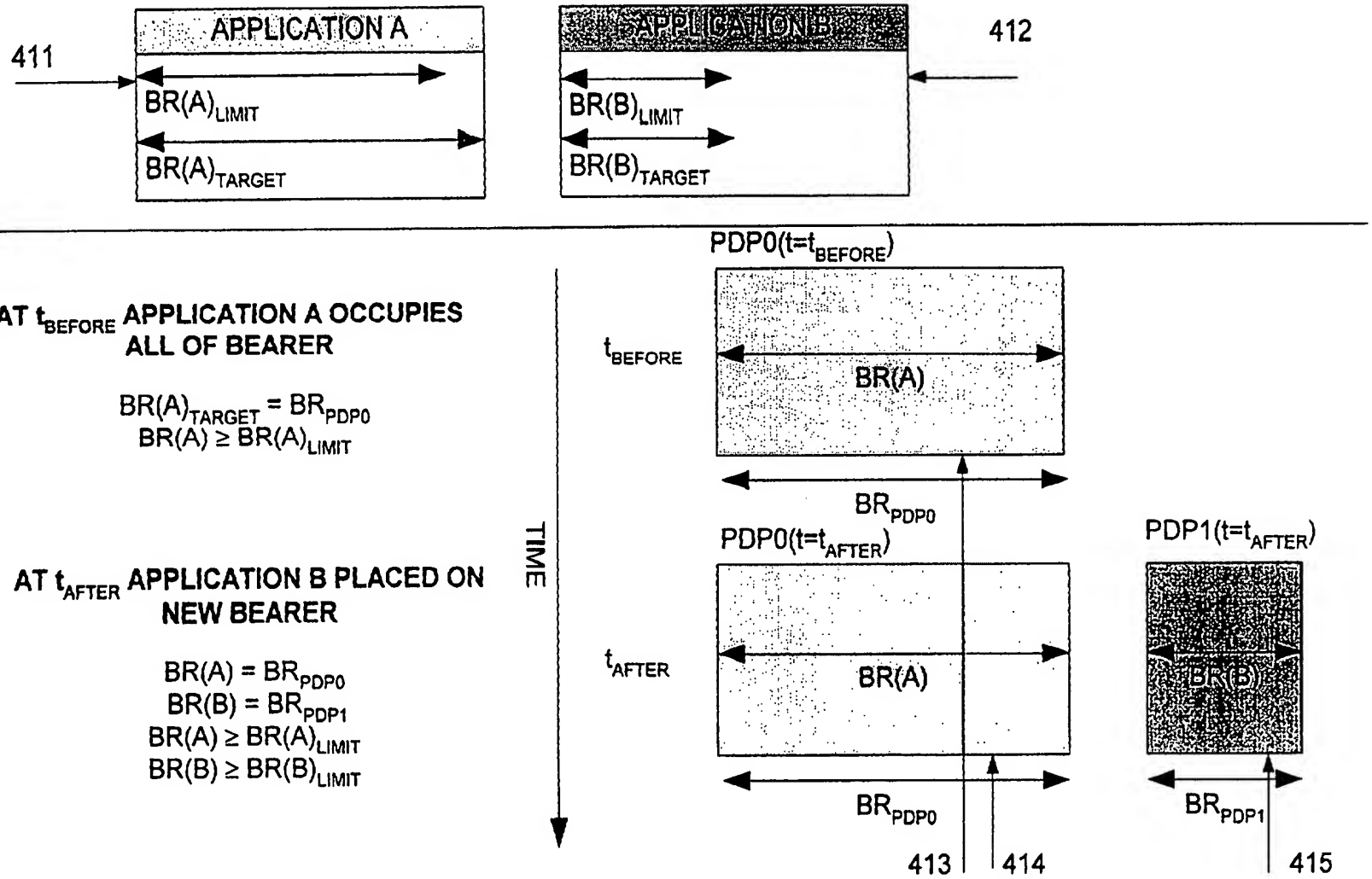

Figure 11

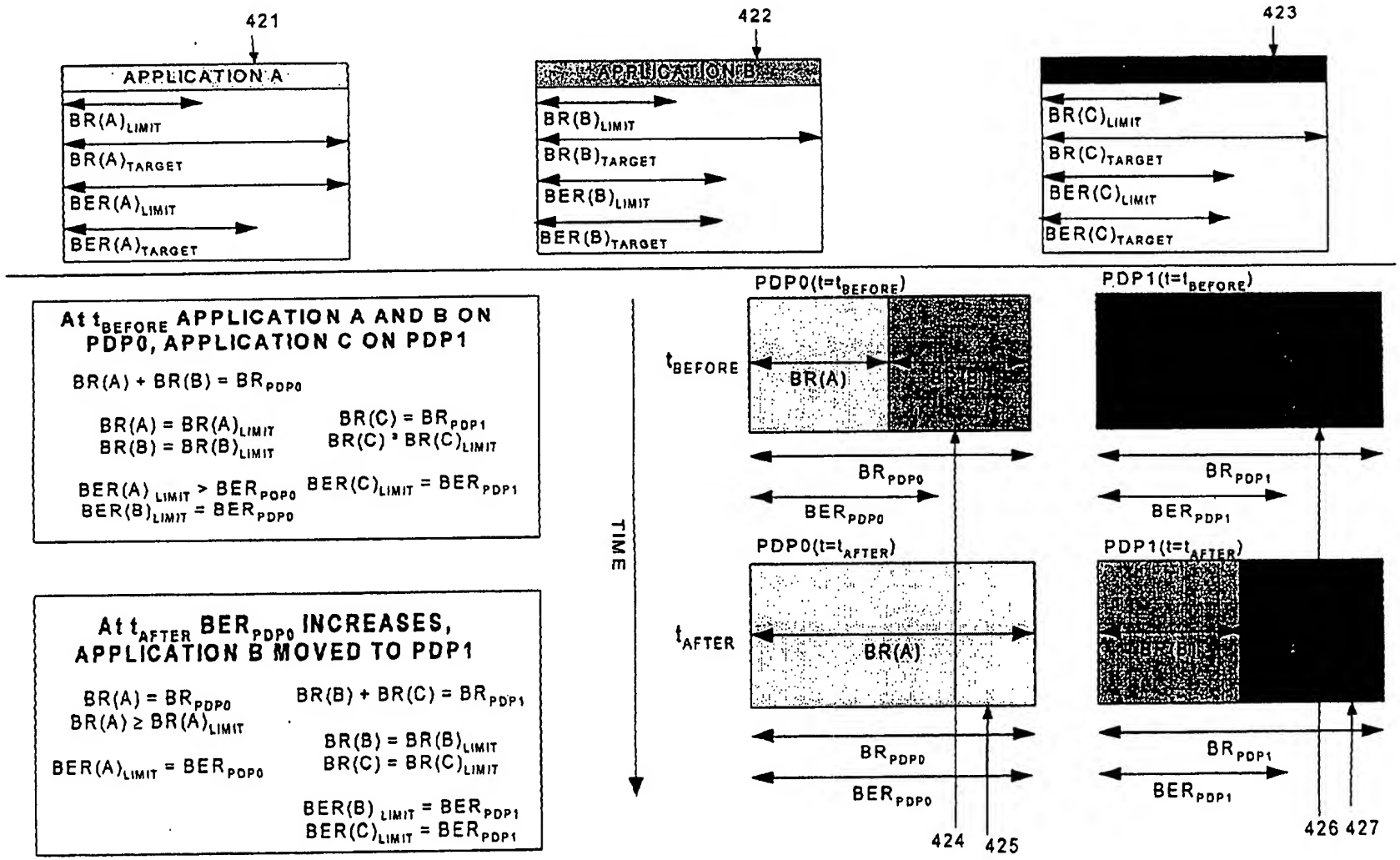

Figure 12

2386282

P103917GB

PACKET DATA COMMUNICATIONS NETWORKS

FIELD OF THE INVENTION

The present invention relates to packet data communications networks, and more particularly, to optimising the allocation of transmission resources between applications executing on a packet data communications device.

The invention has been developed primarily for mobile wireless communications networks such as GPRS and UMTS. However, it will be appreciated by those skilled in the art that the invention is not limited to use with those particular technologies.

BACKGROUND TO THE INVENTION

In UMTS an application can specify the quality of service it requires for each of the data streams that it uses to send and/or receive packetised data to/from other applications. Each data stream is sent over a packet switch connection. Within UMTS the connection is defined by a Packet Data Protocol (PDP) context and is known as a UMTS bearer. The PDP context is used to specify the attributes of the connection and hence the quality of service provided by that connection. At the radio layer, for each PDP context a separate Radio Access Bearer (RAB) is established, with the quality of service attributes as specified in the PDP context. Consequently, if multiple PDP contexts exist simultaneously, then multiple corresponding RABs will also exist to support each UMTS bearer.

The attributes that can be set for each PDP context are as follows:

- Traffic class
- Maximum bit rate
- Guaranteed bit rate
- Delivery order
- Maximum SDU size
- SDU format information
- SDU error ratio

- Delivery of erroneous SDUs
- Transfer delay
- Traffic handling priority
- Allocation retention priority
- Source statistics descriptor

For each one of the four traffic classes, a subset of the other attributes is defined as being applicable to that class. For example in both the conversational and streaming classes, all attributes are applicable with the exception of traffic handling priority.

An application executing on a mobile device will request a connection to be set-up for each one of the data streams it will use. The application's view of the connection is termed a socket. When the application requests a socket to be set-up it will specify the quality of service attributes it requires for the connection such that it is appropriate for the type of data to be transmitted. It is obvious that these could be specified in terms of the PDP context attributes or other suitable set of parameters which can be mapped to PDP context attributes. The entity that deals with the application's connection requests is termed a QoS Manager. It will determine if an existing PDP context could be used for the connection or whether a new UMTS bearer should be set up. Either way, once a UMTS bearer with the appropriate quality of service characteristics has been obtained, the QoS manager will associate the application's socket with that bearer. In this way packets sent on a particular socket are delivered to a particular bearer with the correct QoS attributes.

The network may change the attributes of a UMTS bearer at any time, typically in response to changes in traffic loading. The network notifies the QoS manager when a modification to a PDP context is required, and it will then in turn notify the applications that have sockets that are bound to the changed PDP context. The application can then adjust its data transfer accordingly, or may even decide to terminate if the required quality of service is no longer available. An application may also request a modification to the QoS attributes of any of its connections, which may result in the mobile device initiating a PDP modification to the network.

The above works well if each application has a unique PDP context and hence bearer for each data stream, as the application can respond directly to changes. However, when two or more sockets are bound to a single PDP context and the bit rate of the bearer changes, the QoS manager must determine the impact to each of the affected data streams. The fairest way is to simply adjust the bit rate allocated to each socket by the same fraction, and in the absence of any further information this is all that can be achieved. Some applications will be able to reduce their output accordingly, whilst some applications will terminate their execution, as the quality of service required no longer exists. A more intelligent QoS manager with more information at its disposal could make better decisions about how a change in a PDP context should affect the individual data streams of each application. In the above example, if the QoS manager knew how each application would be affected by a change in the bit rate, then it could adjust the bit rate for each socket such that at least the minimum required quality of service was maintained for both applications.

SUMMARY OF INVENTION

According to a first aspect, the present invention provides a method of optimising the allocation of shared transmission resources between two or more packet data streams of applications executing in a device in a packet data communications network, in which said device includes a transmission resource manager function, the method characterised in that each application specifies a range of acceptable values for one or more of the parameters that determine the transmission quality of its data stream; and the transmission resource manager function allocates the available transmission resources to all the data streams in dependence on the range of acceptable values supplied by each application.

According to a second aspect, the invention provides a device comprising the means for optimising the allocation of shared transmission resources between two or more packet data streams of applications executing on said device in a packet data communications network, said device including a transmission resource manager function, and means for the application to specify a range of acceptable values for one or more of the parameters that determine the transmission quality of its data stream; the transmission resource manager function being adapted to allocate the available transmission resources to all the data streams in dependence on the range of acceptable values supplied by each application.

Preferably, the application supplies for each parameter a range in the form of a target value, which is the preferred value for the application, and a limit value, which is the minimum acceptable value for the application.

In a preferred embodiment, one of the parameters specifies the guaranteed bit rate for a packet data stream.

In a preferred embodiment, one of the parameters specifies the maximum bit rate for a packet data stream.

In a preferred embodiment, one of the parameters specifies the bit error rate for a packet data stream.

In a preferred embodiment, one of the parameters specifies the transfer delay for a packet data stream.

An application can request a new packet data stream from the transmission resource manager function of the device, and the transmission resource manager function preferably attempts to acquire the transmission resources required to provide the target value of each of the transmission parameters specified by the application from the network and when the required resources are not available from the network, reallocates the resources already amongst the data streams of the applications such that the number of data streams utilising the available transmission resources is maximised and each data stream is allocated at least the limit value for each transmission parameter specified; each application being informed by the transmission resource manager function of the new allocation.

The network can change the resources allocated to the transmission resource manager function of a device at any time, in which case the transmission resource manager function preferably reallocates the available resources amongst the data streams of the applications such that the number of data streams utilising the available transmission resources is maximised; and each data stream is allocated at least the limit value for each transmission

parameter specified; and each application is informed by the transmission resource manager of the new allocation.

An application can change the ranges requested for any transmission parameter of any packet data stream at any time, and the transmission resource manager function then preferably attempts to acquire the transmission resources required to provide the target value of each of the transmission parameters specified by the application from the network; and when the required resources are not available from the network, reallocates the resources already acquired amongst the data streams of the applications such that the number of data streams utilising the available transmission resources is maximised, and each data stream is allocated at least the limit value for each transmission parameter specified; each application being informed by the transmission resource manager of the new allocation.

In a preferred embodiment, the transmission resource manager function determines the actual value of the parameters of the transmission resources to request from the network for a packet data stream ensure that the value to be requested for a parameter is within the range specified by the application for said packet data stream; and the transmission resource manager function uses tariff information for determining the cost per unit time of the transmission resources at the current time of day, and ensures that the cost per unit time of the transmission resources to be requested from the network for all the packet data streams of an application or all the packet data streams active on the device are within a user specified range.

In a preferred embodiment, the device is a mobile terminal attached to a GPRS packet data network.

BRIEF DESCRIPTION OF DRAWINGS

Preferred embodiments of the invention will now be described, by way of example only, with reference to the accompanying drawings, in which:

Figure 1 is a simplified schematic diagram of a packet data communication system showing terminals, an access network, a core network, and other packet data networks;

Figure 2 is a simplified schematic diagram of a UMTS communication system showing two mobile terminals, a radio access network, a UMTS backbone and a gateway to other IP networks;

Figure 3 is a simplified schematic of the architecture of the protocol stack of a wireless packet data communications device that supports QoS;

Figure 4 shows the different types of PDP QoS attributes for a RAB;

Figure 5 is an entity relationship diagram in UML notation showing the relationships and cardinalities of the relationships between objects involved in quality of service enabled packet data communications;

Figure 6 is a message sequence chart showing the sequence of events and actions when an application running on a UMTS mobile terminal opens a QoS enabled socket;

Figure 7 is a message sequence chart showing the sequence of events and actions when a UMTS network initiates a change in the quality of service attributes of a UMTS bearer in use by a UMTS mobile terminal;

Figure 8 is a message sequence chart showing the sequence of events and actions when a UMTS network reduces the bandwidth of a RAB over which an application is transmitting on the uplink;

Figure 9 is a message sequence chart showing the sequence of events and actions when a UMTS network reduces the bandwidth of a RAB over which an application is receiving on the downlink;

Figure 10 is a diagram illustrating how the bandwidth or a wireless bearer is shared between a first and a second application, when the first application has a flexible QoS policy profile;

Figure 11 is a diagram illustrating how the bandwidth or a wireless bearer is not shared between a first and a second application, when the first application has a inflexible quality of service policy profile, resulting in the establishment of a second bearer for the second application; and

Figure 12 is a diagram illustrating how a network initiated change in the bit error rate of one wireless bearer can result in a reconfiguration of the mapping of applications to wireless bearers and a corresponding change in the bandwidth allocated to the applications, in order to accommodate all applications within the quality of service attribute ranges specified in their quality of service policy profiles.

DETAILED DESCRIPTION OF PREFERRED EMBODIMENTS

The preferred embodiment of the present invention is applied to packet data communication terminals that are able to provide guarantees of quality of service (QoS) of packet data communications to applications executing on the terminal and the associated networks that support guaranteed QoS packet data communication.

Referring to the drawings, and in particular Figure 1, there is shown a packet data communications network 100. The network is composed of different network elements, which cooperate to provide the end-to-end connectivity. Of these elements, the packet data terminal 101 is the endpoint of communication on the network. It provides the initial point of access to the network for user applications. Each packet data terminal has at least one network address that uniquely identifies it to the network and hence allows packetised data to be sent to a specific terminal based upon this address. The packet data terminal is physically connected to the access network 105 of the packet data communications network

by the access link 103. The access network 105 and the access links 103 may be either wireline or wireless.

The access network 105 manages terminal connectivity with its access control function 106. This function determines whether terminals are authorised to connect, the resources they are allowed for connection based upon policy, and provides accounting of the usage of terminal or access resources. It is also responsible for the resource management of the access network and providing guarantees on QoS of access links used by terminals. The access network 105 is connected by a high bit rate connection 107 to the core packet data network. Thus, one aspect of the access network is to convert between the link layer technologies of the access link to the link layer technology of the core packet data network.

The core packet data network provides connectivity to a plurality of access networks and hence packet data terminals. The core packet data network also provides connectivity 109 to other packet data networks 110, and hence to packet data terminals in other networks. The core packet data network may be QoS enabled in that it provides guarantees on, for example, the bandwidth available or the delay for a packet data stream, between two packet data terminals. It will be appreciated that a single core packet data network may have a plurality of access networks that in turn provide access to a plurality of packet data terminals. The core packet data network may also provide connectivity to a plurality of other packet data networks.

Two terminals on a packet data network communicate using a packet data protocol such as Internet Protocol version 4 (IPv4), and Internet Protocol version 6 (IPv6). This protocol prescribes how the packets are constructed and the addressing of network elements. Packets are routed between terminals of the network based upon the network address by the packet routers of the access network and the core packet data network. The packet data protocol provides independence of the underlying link layer technology such that

applications requiring packet data communications need only understand the packet data protocol.

It will be appreciated that Figure 1 and the above description is a schematic view of a packet data communications network, and that many different installations are possible, utilising a wide range of technologies and protocols. A specific example is shown in Figure 2, which is a Universal Mobile Telecommunications System (UMTS) packet data network. The UMTS packet data network will be used throughout but the invention is not restricted to a UMTS packet data network.

The mobile terminals (MT) 121 are the endpoints of communication on the network. They are connected to the Radio Access Network (RAN) 127 via the Uu interface, which uses Wideband Code Division Multiple Access (WCDMA) technology. The RAN is composed of the so-called Node-B base stations, and the Radio Network Controller (RNC) 129 which manages the resources of the radio network. The core packet data network is effectively composed of the Serving GPRS Support Node (SGSN) 131, which provides connectivity to access networks, the UMTS backbone 133, and the Gateway GPRS Support Node (GGSN) 135, which provides interconnectivity to other IP networks 137. The UMTS backbone provides connectivity between SGSN and GGSN. The connection between MT and SGSN is known as the Radio Access Bearer (RAB) Service. The QoS of each Radio Bearer that is set up is managed by the RNC. It also performs dynamic allocation of the finite radio resources available for data transmission across the RABs from the terminals.

QoS for packet data communication means guaranteeing that the characteristics of an end-to-end application data connection will be maintained for said application data connection. These characteristics include, but are not limited to, bit rate, delay, and error rates. In order to achieve these guarantees, the network must ensure that:

- data packets carry an indication of how they should be treated by the network; and
- resources are reserved across the network so that phenomena such as congestion are avoided; and

- the links of the network path that a packet is sent on have the desired characteristics such as bit error rate.

For IP networks the Internet Engineering Task Force (IETF) defined protocols DIFFSERV, INTSERV and RSVP perform these functions.

Referring now to Figure 3, there is shown a packet data communications terminal 300. Some operational units and protocol layers of the terminal are shown where required for the description of the invention. All possible operational units and protocol layers are not shown. It will also be appreciated that there could be many specific instances of protocol layers and operational units used in the terminal, where in Figure 3, a schematic of the protocol layer or operational unit is used. The terminal provides an execution environment for applications 311. Examples of applications which require packet data communications with other terminals or servers include, but are not limited to, email (using SMTP), web browsers (using HTTP), packetised voice (using SIP and RTP/RTCP) and streaming video (using RTSP and RTP/RTCP). Applications 311 which execute on the terminal 300 are provided access to packet data communications by the terminal's operating system using a socket application programming interface (Socket API) 310. This interface provides programmatic routines that allow an application to open a data stream for reading and/or writing between itself and another application resident on another terminal or server in the network. A second API 309 is provided to the application by the terminal's operating system through which the application can request that particular QoS attributes are guaranteed for a data stream. This is termed the quality of service application programming interface (QoS API). It is known that the Socket API and the QoS API may be combined into a single application programming interface or they maybe be separate APIs from the applications perspective. Using the QoS API the application can request that characteristics of the data stream such as the mean bandwidth, peak bandwidth, delay, error rate, etc. are set in accordance with the applications needs. For example, a video telephony application requires that the data streams it uses have a very low delay and hence would request low delay for each data stream through the QoS API.

The packet data protocol layer 308 is an implementation of packet data communications protocol on the terminal and is accessed by the application through the Socket API. There are many different packet data protocols such as Internet Protocol version 4 (IPv4), Internet Protocol version 6 (IPv6) and X.25. This protocol layer takes in data from the application via a socket and processes and formats it in accordance with the protocol. The protocol layer can provide various different types of packet data communication to the application. For example, IPv4 and IPv6 provide an unreliable datagram service using UDP and a reliable streaming service using TCP. The packet data protocol layer may also provide mechanisms for guaranteeing end-to-end quality of service. For example in IPv4 RSVP can be used to reserve resources across a network.

The line 301 divides the packet data protocol layer 308 and the link layer protocols. The link layer protocol (303 and 304) is responsible for establishing channels in the underlying physical transmission medium and for formatting and processing the packets received from the packet data protocol layer for transmission. The link layer protocol is separated into two parts, a control function 303 and a data function 304. The control function 303 is used for establishing and managing channels, setting up QoS characteristics for the channel and responding to network initiated changes in the QoS. The data function 304 performs the task of processing the packets received from the packet data protocol layer and transmitting them on to the actual physical network interface, and processing packets received from the physical network interface and transmitting them to the packet data protocol layer.

Between the packet data protocol layer and the link layer protocol sits the packet classifier 307. Its function is to map packets from the packet data protocol layer 308 onto the channels provided by the data transmission function 304. In order to do this, it needs to know from which data stream a packet originated and which data transmission channel is associated with that socket. This information is provided by the quality of service manager (QoSM) 305 over the interface 302.

The QoSM 305 provides the implementation behind the QoS API. In the prior art it is known to provide the following functions:

- determine whether a data transmission channel needs to be established or whether an existing data transmission channel is to be reused in response to an application request for a specific QoS for a data stream;
- request the establishment of data transmission channels with the application specified QoS characteristics from the link layer control function 303;
- negotiate with the network the acceptable QoS characteristics for a data transmission channel;
- inform the application of the QoS assigned to a data stream;
- receive from the network any changes to the QoS characteristics of a data transmission channel;
- receive from the application requests for changes to the QoS characteristics of a data stream;
- inform the application of network initiated changes in QoS for a data stream;
- interact with the packet data protocol in establishing end-to-end QoS guarantees;
- inform the packet classifier of the mapping of data streams to data transmission channels;
- dynamically manage the association of data streams and data transmission channels such that the closest match between the application requested QoS and the available QoS in the network on a given data transmission channel is achieved at all times;
- assigning default QoS attributes for data streams when not specified by the application.

It uses the interface 302 to control the packet data protocol and the packet classifier.

The QoS API provides a way for the application to request that a particular data stream has certain attributes, which are collectively termed the QoS Profile. In a UMTS packet data network, each bearer has values for each of these attributes:

- Traffic class
- Uplink maximum bit rate, Downlink maximum bit rate
- Uplink guaranteed bit rate, Downlink guaranteed bit rate
- Delivery order

- Maximum SDU size
- SDU format information
- SDU error ratio
- Residual bit error ratio
- Delivery of erroneous SDUs
- Uplink transfer delay, Downlink transfer delay
- Traffic handling priority
- Allocation retention priority
- Source statistics descriptor

The specific set of values of the QoS attributes associated with a bearer is contained within the PDP Context. The actual values of the QoS attributes of each PDP context are controlled by the network and may be modified by the network at any time. The network manages the available resources such that each user is allocated a fair share in accordance with the network operator's policy. The application may also request at any time a modification to the QoS Profile associated with a particular data stream. The QoS attributes of the PDP context can be categorised as either:

- asymmetric cumulative; or
- asymmetric non-cumulative; or
- symmetric non-cumulative.

Asymmetric attributes have independent values in the send and receive directions respectively, and they may be the same or they may be different. Symmetric attributes must take on the same value in both the send and receive directions. Non-cumulative attributes have the property that every data stream that is using the PDP context takes on the same value for that attribute. For example, the residual bit error ratio is non-cumulative as it is the same for all data streams on the same bearer. Cumulative attributes are those attributes that can be must be sub-divided between the data streams using the bearer. For example, if three data streams are sharing a single PDP context, then the guaranteed uplink bit rate of the PDP context must be shared between the data streams. The value of cumulative PDP context attributes can be changed, but this does not necessarily result in a change of said

attribute for all application data streams using that PDP context. Figure 4 shows the breakdown of the PDP QoS attributes.

Referring now to Figure 5 there is shown a relationship diagram in the Unified Modelling Language (UML) notation, which illustrates the relationship between the key entities in a packet data terminal involved in IP packet data communication with QoS. An application 208 executing on a packet data terminal opens sockets to write 202 and sockets to read 322 for its transmit data streams 201 and receive data streams 321. An application can open one or more read or transmit data streams and hence read or write sockets, but only one application can use any specific socket. The operating system of the packet data terminal associates each write socket with a physical bearer downlink 203 which will be used for the transmission of packet data, and each read socket with a physical bearer uplink 323 which will be used for the reception of packet data. A single physical bearer 207 may be used to transmit or receive the data packets of one or more sockets.

An IP address identifies an endpoint of communication to the network. Each data packet contains a recipient address, which is composed of an IP address and a port number so that any packet can be first delivered to the correct terminal and secondly delivered to the correct socket and hence application on said terminal. The combination of IP address and port number is known as the transport address. Associated with each write socket 202 are a destination IP address of a remote host 205 and a port number on that remote host. Associated with each read socket 322 are an IP address on the local host 325 and a port number on the local host.

Each physical bearer 207 has a set of bearer QoS attributes 206 which defines the QoS that packets transported on said bearer will receive on the uplink and downlink. For UMTS packet data this is known as the PDP context. When the application opens a socket for reading or writing it will specify the QoS it requires for that socket. The QoS manager in the packet data terminal must associate with that socket a physical bearer with the same QoS attributes specified by the application for the socket. Once the resources have been allocated to the socket, the QoS manager informs the application of the values of the QoS attributes allocated to the socket.

Referring now to Figure 6 there is shown a message sequence chart of the prior art for an application 210 executing on a UMTS mobile terminal 219 opening a QoS enabled IP data stream for communication with a server 215 in the packet network. Firstly, the application 210 requests the QoS Manager 211 to create a connection with the required QoS attributes in the UMTS network 221. This triggers the QoS manager to request the UMTS protocol stack 213 in the MT to establish a PDP context whose QoS attributes correspond to those requested by the application 222. The UMTS protocol stack creates an Activate PDP Context Request 223 and sends it to the UMTS network 214. The UMTS network responds with a positive acknowledgement 224 and the PDP context is established in the network. This acknowledgement is received by the UMTS stack, which forwards it 225 to the QoS manager. With the PDP context now set up and hence also the UMTS bearer, the QoS manager can now open the end to end IP connection with the server through the IP stack 226. The details of the TCP/IP protocol exchange 227 are not shown. When completed, the IP stack acknowledges positively to the QoS manager 228, which then provides the application with a reference to the socket 229. End-to-end QoS can be established by the QoS manager once the PDP context and associated RAB has been established, but before returning the socket reference to the application. Further detailed information is available in the Third Generation Partnership Project (3GPP) technical specifications TS23.107 and TS23.207.

Referring now to Figure 7 there is shown a message sequence chart of the known art for a network-initiated change in the QoS of a UMTS bearer. The SGSN 121 detects a need to modify the QoS attributes of a PDP context 231, possibly due to cell loading for example. It instructs 232 the Radio Access Network 127 to modify the Radio Access Bearer in accordance with the new QoS attributes supplied 233. Once the RAN has completed and informed the SGSN 234, the SGSN requests the GGSN that the PDP context of the UMTS bearer is updated 235. The GGSN updates its internal memory 236, and then signals the change 237, 238 to the Mobile Terminal 219 via the SGSN. The QoS Manager 211 in the MT handles the PDP context modification message and updates its internal memory of the PDP context for that bearer to reflect the change in QoS 239. It acknowledges the PDP context modification back to the network 241. The QoS manager then determines which applications have data streams that are affected by the PDP context change and informs

them of the new settings in turn 242, 243, 244. Finally, if any application's data streams were assigned to a new PDP context, then the Traffic Flow Template in the GGSN must be updated 245.

The prior art applications can specify the attributes that they require of the data transmission path to achieve a certain QoS, and the QoSM will allocate the resources if available or request more resources from the network. In the prior art, it is only known for applications to specify a single value of each QoS attribute that they require for each data stream. For example, in a UMTS packet data network they can specify a residual bit error ratio or a maximum transmit bit rate or the required guaranteed bit rate. This enables the QoSM to match the data stream to the physical bearer with the desired characteristics. It may happen that:

- an existing PDP context has matching QoS attributes and is reused. The extra bandwidth required to accommodate the new data stream on said PDP context is requested from the network;
- there is no existing PDP context with matching QoS attributes and the QoSM requests a new PDP context and hence physical bearer is setup by the network;

In either case, the request for resources from the network may fail, and the application must be informed that its QoS requirements cannot be met. It is then up to the application to determine what it should do next. It may choose to terminate or request a lower QoS for its data stream.

An application can request at any time that its QoS for a data stream should be changed. In a first example, the application using an uplink makes a request of the QoSM for more bandwidth. In this case, the QoSM can request that the network allocated more bandwidth to the PDP context. This may be successful, but it may also fail. If it fails, then the QoSM may have the option of reallocating bandwidth that other applications are using on the same PDP context. However, the QoSM does not know how a reallocation of bandwidth on the uplink will affect the transmit data streams of other applications. In a second example, the application using an uplink may request a reduction in the bandwidth it requires for a

transmit data stream. In this case, the QoSM can simply return the resources to the network, or alternatively reallocate it to the data streams of other applications using the uplink that are active. However, again the QoSM cannot make an intelligent decision about how to allocate the excess bandwidth to applications using the uplink.

The network can change the PDP context at any time in accordance with its resource control algorithms, which affects the QoS granted to the data streams whose sockets are using the uplink or downlink of the corresponding physical bearer. In a first example, the RNC reduces the uplink bandwidth of a PDP context. The QoSM must decide how each transmit data stream using the uplink is affected by this change, and inform the affected applications. In the case where there is only a single transmit data stream on the PDP context, this is straightforward. The QoSM must pass the entire reduction on to the single transmit data stream and it's associated application. However, when there are two or more transmit data streams the QoSM must distribute the reduction on to each data stream and associated application in equal measures, because it does not know how each transmit data stream and associated application will be affected. This is not necessarily the optimal action, as one application may be more tolerant to a reduction in bit rate than the other, but the QoSM does not have this information. Alternatively, one application may accept the same bit rate on a data stream but at a higher bit error rate, which may be available.

In a second example, the radio network controller increases the bandwidth of a PDP context. As in the first example, the QoSM lacks the information to make an intelligent decision about how this new bandwidth should be optimally distributed between the data streams and their respective applications sharing this PDP context.

It is a feature of the present invention that the application makes available to the QoSM additional information about the QoS it requires for each of its data streams, such that the QoSM is able to manage the resources made available by the network between multiple applications more effectively than in the prior art.

In a preferred embodiment, the application provides to the QoSM what is termed a QoS policy profile (QPP) for each socket that it opens. This extends the apparatus and method

of the QoS API, in that it allows an application to specify for one or more of its data streams information for each QoS attribute that specifies:

- The value that the application requires for ideal operation. This is termed the target value.
- The value beyond which the application cannot function. This is termed the limit value.

In a preferred embodiment the following QoS attributes would be specified in terms of target, and limit values:

- Maximum bit rate
- Guaranteed bit rate
- SDU error ratio
- Transfer delay

It will be appreciated by those skilled in the art that the target and limit values can be applied to any QoS attribute that can take on a range of values and it is not limited to those specified above.

In a preferred embodiment, the user can change the QPP for an applications packet data streams through an application configuration tool. In another embodiment, applications will have these values permanently fixed by the application developer.

The QoSM utilises the QPP supplied by the application for the data streams to perform intelligent management of the transmission resources allocated to it by the network such that:

- Applications do not monopolise scarce resources unnecessarily.
- Each application can be allocated a fair share of the resources available.
- The maximum possible number of simultaneous streams is maintained.

It is an object of the present invention that the QoSM will always attempt to obtain from the network the target value of a QoS attribute for a data stream, specified by the application in the QPP. If the target values for QoS cannot be obtained, for example, because the network is heavily loaded, then the QoSM will reallocate some of the available resources such that each data stream of each application has at least the limit values specified in its QPP for its allocated QoS. If the available resources are limited such that there are not enough resources to satisfy the limit values of all the QPPs then the QoSM will use an algorithm to determine which QPPs will be satisfied with their limit resources and which will not be satisfied. This algorithm can be, though is not limited to, one of the following:

- Last in first out allocation of resources
- Maintaining the maximum possible number of open data streams.

If the available resources are increased, the QoSM will allocate them according to an algorithm such as, though not limited to:

- The maximum possible number of data streams to meet their QPP targets for the maximum possible number of values;
- An allocation based upon the difference between current value and target values;
- An allocation based upon the ratio of current value and target values.

The QoSM will use the QPP as described above whenever a change in resource allocation is required, specifically:

- When an application requests the opening of a new data stream with a new QPP;
- When an application closes an existing data stream;
- When an application modifies an existing QPP;
- When the network increases the allocated resources;
- When the network decreases the allocated resources.

According to another aspect of the invention, the QoSM will use tariff information available to it, for example, in a terminal's memory, which provides information about the costs of different QoS at different times of the day, along with a specified user preference for how much a packet data communication session for the specific application should cost, to determine whether

- the target value from the QPP should be requested;
- the limit value from the QPP should be requested;
- a value in-between the target and limit should be requested.

Thus the QPP enables the QoSM to vary the QoS provided to an application in order to maintain a user specified cost for the service.

The invention is equally applicable whether an application is transmitting data, receiving data or both over the UMTS network. Transmit data streams will be transmitted over a RAB uplink, and receive data streams will be received over a RAB downlink. For each type of data stream, a QPP will be specified by the application to the QoSM. In the case of transmit data streams, the application will set the target and limit of QoS parameters based on its requirements. In the case of receive data streams, it will set the target and limit of QoS parameters based on the requirements of the entity transmitting. These requirements will have been communicated to the application from the entity transmitting.

If after the QPPs have been setup, and the application is receiving or transmitting data, the PDP context of a RAB is changed by the network, the QoSM takes action using the QPPs and notifies the application. In the case of transmit data streams the application will use the modified QoS parameters received from the QoSM to modify its output as in Figure 8. An application 500 is transmitting data 510 to a receiver 502 over a RAB. At some time later, a network generated reduction in the bandwidth of this bearer occurs 511. Through a dialog with the QoSM the application is notified of this reduction and reduces it's output bitrate 512.

In the case of receive data streams, the application will pass on notice of these changes in the QoS parameters to the transmitting entity as in Figure 9. An application 500 is receiving data 520 from a transmitter 503. At some time later, a network generated reduction in the bandwidth of this bearer occurs 511. Through a dialog with the QoSM the application is notified of this reduction and notifies the transmitter 522. Transmitter 503 then reduces it's output bitrate 523. Also as a result of the reduction of bandwidth 511, it may be necessary to reallocate the receive data streams onto different RABs. This would involve updating the Traffic Flow Templates (TFT) of the existing PDP contexts.

As illustrated in Figure 10, an application 401 has one transmit data stream using a PDP context PDP0($t=t_{BEFORE}$) 403. Its target bit rate, $BR(A)_{TARGET}$, is equal to the uplink guaranteed bit rate of PDP0($t=t_{BEFORE}$) 403, BR_{PDP0}. Its limit bit rate, $BR(A)_{LIMIT}$, is half of its target bit rate, $BR(A)_{TARGET}$. Another application 402 is launched which can use the same PDP context (because QoS attributes such as error rate or delay are suitable for the second application) PDP0($t=t_{AFTER}$) 404 (may even be the only one available as with GPRS) for its transmit data stream. No more resources are available from the network. The QoSM can then determine that both applications can share the PDP0($t=t_{AFTER}$) 404 if the sum of their limit bit rates, $BR(A)_{LIMIT} + BR(B)_{LIMIT}$, is less than the PDP bit rate, BR_{PDP0}. Each transmit data stream is assigned a portion of the bit rate such that the available bandwidth on the PDP0($t=t_{AFTER}$) 404 is fully utilised. Without the QPP, the application must either request the minimum values acceptable, e.g. BR_{LIMIT}, or request its ideal value, e.g. BR_{TARGET}. In the first case, the application will always deliver to the user a sub-optimal experience, but will not monopolise the available resources. In the second case, it can deliver to the user high quality but will monopolise the resources and starve other applications to the extent that it can prevent them from executing.

As illustrated in Figure 11, an application 411 has one transmit data stream using a PDP context PDP0($t=t_{BEFORE}$) 413. Its target bit rate, $BR(A)_{TARGET}$, is equal to the uplink guaranteed bit rate of the PDP0($t=t_{BEFORE}$) 413, BR_{PDP0}. Its limit bit rate $BR(A)_{LIMIT}$ is, for example, 80% of its target bit rate, $BR(A)_{TARGET}$. At some time later, t_{AFTER}, another application 412 is launched which can use the same PDP context PDP0($t=t_{AFTER}$) 414 for its transmit data stream as the QoS attributes match. However, the sum of the two limit bit

rates, $BR(A)_{LIMIT} + BR(B)_{LIMIT}$, is greater than the total available bandwidth on the uplink of the bearer BR_{PDP0}. The QoSM then is able to determine that it must establish a new PDP context $PDP1(t=t_{AFTER})$ 415 for the second application 412. The original PDP context PDP0 remains unchanged i.e. $PDP0(t=t_{BEFORE}) = PDP0(t=t_{AFTER})$.

Consider now the situation illustration in Figure 12 in which two transmit data streams belonging to different applications 421 and 422 share a PDP context $PDP0(t=t_{BEFORE})$ 424. Both transmit data streams have only their limit value of the bit rate allocated. The bit error rate (BER) of the $PDP0(t=t_{BEFORE})$ 424, BER_{PDP0}, is the limit value for one transmit data stream ($BER(B)_{LIMIT} = BER_{PDP0}$), but is the target value for the other ($BER(A)_{TARGET} = BER_{PDP0}$). Note that for BER, a lower value represents superior quality. Another transmit data stream belonging to a third application 423 is mapped to a separate PDP context $PDP1(t=t_{BEFORE})$ 426 with the same BER as that for $PDP0(t=t_{BEFORE})$ 424, i.e. $BER_{PDP0} = BER_{PDP1}$. Now at some later time, t_{AFTER}, the network increases BER_{PDP0}. Because the QoSM knows that one of the transmit data streams on $PDP0(t=t_{AFTER})$ cannot accept a higher BER, i.e. the transmit data stream belonging to application 422, it moves that transmit data stream to $PDP1(t=t_{AFTER})$ where bandwidth is made available by reducing that available to application 423, since on $PDP1(t=t_{AFTER})$ 427 the BER is lower than on $PDP0(t=t_{AFTER})$ 425. The application 421 is then able to utilise all the bandwidth on $PDP0(t=t_{AFTER})$. This is only possible when a range of acceptable values for a QoS attribute is provided. Without this information, the QoSM would have increased the BER_{RAB0} which would have terminated application 422 which could not tolerate this.

The cost of network resources varies over time. Peak time usage is more expensive than off-peak. The QoSM can use this tariff information combined with the QPP to trade off cost and quality in accordance with the users configuration.

The usage of a QPP is independent of where it is stored during application execution. In the preferred embodiment, it is passed to the QoSM, which can then autonomously make decisions about how to manage the QoS. In another embodiment, it is retained in the application, and the QoSM will offer new QoS attributes for a data stream to the application, which the application can then either accept or reject based upon its QPP. This

embodiment is less efficient because of this negotiation procedure between the applications and the QoSM. It is also only possible for QoS- aware applications.

The characteristic feature of this invention is that specifying a limit value and a desired value for each QoS attribute enables the QoSM to perform intelligent management of the transmission resources.

CLAIMS

1. A method of optimising the allocation of shared transmission resources between two or more packet data streams of applications executing in a device in a packet data communications network, in which said device includes a transmission resource manager function, the method being characterised in that:

   each application specifies a range of acceptable values for one or more of the parameters that determine the transmission quality of its data stream; and

   the transmission resource manager function allocates the available transmission resources to all the data streams in dependence on the range of acceptable values supplied by each application.

2. A method according to claim 1, wherein the application supplies for each parameter a range in the form of a target value, which is the preferred value for the application, and a limit value, which is the minimum acceptable value to the application.

3. A method according to any preceding claim, wherein one of the parameters specifies the guaranteed bit rate for a packet data stream.

4. A method according to any preceding claim, wherein one of the parameters specifies the maximum bit rate for a packet data stream.

5. A method according to any preceding claim, wherein one of the parameters specifies the bit error rate for a packet data stream.

6. A method according to any preceding claim, wherein one of the parameters specifies the transfer delay for a packet data stream.

7. A method according to any preceding claim, wherein an application requests an increase in transmission resources from the transmission resource manager function;

   the transmission resource manager function reallocates the resources already acquired amongst the data streams of the applications such that the number of data streams utilising the available transmission resources is maximised and each data

stream is allocated at least the limit value for each transmission parameter specified; and

each application is informed of the new allocation for each affected packet data stream.

8. A method according to claim 7 in which the transmission resource manager function responds to a request for increased transmission resources by attempting to acquire the transmission resources required to provide the target value of each of the transmission parameters specified by the application from the network; and if the required resources are not made available by the network, reallocating the resources already acquired amongst the data streams of the applications.

9. A method according to claim 7 or 8 wherein an application requests a new packet data stream.

10. A method according to claim 7 or 8, wherein an application changes the range requested for any transmission parameter of any packet data stream so as to require increased transmission resources.

11. A method according to any preceding claim, wherein an application reduces its requirement for transmission resources compared with those previously acquired by the transmission resource manager function; the transmission resource manager function reallocating the required resources amongst the data streams of the applications and freeing up any remaining network resources such that the number of data streams utilising the available transmission resources is maximised and each data stream is allocated at least the limit value for each transmission parameter specified; and

each application is informed of the new allocation for each affected packet data stream.

12. A method according to any preceding claim, wherein the network changes the resources allocated to [and controlled by the transmission resource manager function of] a device at any time;

the transmission resource manager function reallocates the available resources amongst the data streams of the applications such that

the number of data streams utilising the available transmission resources is maximised and each data stream is allocated at least the limit value for each transmission parameter specified; and

each application is informed of the new allocation for each affected packet data stream.

13. A method according to any preceding claim, wherein said transmission resource manager function determines the actual value of the parameters of the transmission resources to request from the network for a packet data stream, the method including the steps of:

ensuring that the value to request for a parameter is within the range specified by the application for said packet data stream; and

using tariff information available to the transmission resource manager function for determining the cost per unit time of the transmission resources at the current time of day; and

ensuring that the cost per unit time of the transmission resources to be requested from the network for all the packet data streams of an application or all the packet data streams active on the device are within a user specified range.

14. A data communications device comprising the means for optimising the allocation of shared transmission resources between two or more packet data streams of applications executing on said device in a packet data communications network, said device including a transmission resource manager, characterised in that the device further comprises means for the application to specify a range of acceptable values for one or more of the parameters that determine the transmission quality of its data stream; and the transmission resource manager is adapted to allocate the available transmission resources to all the data streams in dependence on the range of acceptable values supplied by each application.

15. A data communications device according to claim 14, in which said means supplies for each transmission parameter a range in the form of a target value, which is the preferred value for the application, and a limit value, which is the minimum acceptable value to the application.

16. A data communications device according to claims 14 or 15, in which said means specifies the guaranteed bit rate for a packet data stream.

17. A data communications device according to any one of claims 14 to 16, in which said means specifies the maximum bit rate for a packet data stream.

18. A data communications device according to any one of claims 14 to 17, in which said means specifies the bit error rate for a packet data stream.

19. A data communications device according to any one of claims 14 to 18, in which said means specifies the transfer delay for a packet data stream.

20. A data communications device according to any one of claims 14 to 19, in which said means can request an increase in transmission resources; and in which the transmission resource manager function reallocates the resources already acquired amongst the data streams of the applications such that the number of data streams utilising the available transmission resources is maximised and each data stream is allocated at least the limit value for each transmission parameter specified; and each application is informed of the new allocation for each affected packet data stream.

21. A data communications device as claimed in claim 20 in which the transmission resource manager function responds to a request for increased transmission resources by attempting to acquire the transmission resources required to provide the target value of each of the transmission parameters specified by the application from the network; and if the required resources are not made available by the network, reallocating the resources already acquired amongst the data streams of the applications.

22. A data communications device as claimed in claim 20 or 21 wherein an application can request a new packet data stream.

23. A data communications device as claimed in claim 20 or 21 wherein an application can change the range requested for any transmission parameter of any packet data stream, so as to require increased transmission resources.

24. A data communications device as claimed in any one of claims 14 to 23 wherein an application can reduce its requirement for transmission resources compared with those previously acquired by the transmission resource manager function; the transmission resource manager reallocating the required resources amongst the data streams of the applications and freeing up any remaining network resources such that the number of data streams utilising the available transmission resources is maximised and each data stream is allocated at least the limit value for each transmission parameter specified; and each application being informed by the transmission resource manager of the new allocation for each affected packet data stream.

25. A data communications device wherein when the network changes the resources allocated to and controlled by the transmission resource manager of; the transmission resource manager reallocates the available resources amongst the data streams of the applications such that the number of data streams utilising the available transmission resources is maximised and each data stream is allocated at least the limit value for each transmission parameter specified; and each application is informed of the new allocation for each affected packet data stream.

26. A data communications device according to any one of claims 14 to 25, in which the transmission resource manager determines the actual value of the parameters of the transmission resources to request from the network for a packet data stream; ensures that the value to request for a parameter is within the range specified by the application for said packet data stream; and using tariff information available determines the cost per unit time of the transmission resources at the current time of day; and

ensures that the cost per unit time of the transmission resources to be requested

from the network for all the packet data streams of an application or all the packet data streams active on the device are within a user specified range.

27. A data communications device according to any one of the claims 14 to 26, that is a mobile terminal in a UMTS packet data network.

28. A data communications device according to any one of the claims 14 to 27, that is a mobile terminal in a GPRS packet data network.

Application No:	GB 0205105.0	**Examiner:**	Jared Stokes
Claims searched:	1 to 28	**Date of search:**	25 March 2002

Patents Act 1977
Search Report under Section 17

Databases searched:

UK Patent Office collections, including GB, EP, WO & US patent specifications, in:

UK Cl (Ed.T): H4K (KOT, KOD8, KTKA, KTKX)
H4L (LDGP, LRRMW, LRRMS)

Int Cl (Ed.7): H04L (12/56, 29/06)
H04Q (7/22, 11/04)

Other: On-Line - EPODOC, JAPIO, WPI

Documents considered to be relevant:

Category	Identity of document and relevant passage		Relevant to claims
A	EP 0 734 195 A2	(AT&T) See column 3 line 8-column 5 line 2	-
A	WO 01/89234 A2	(Ericsson) See abstract	-
A	US 5 892 754	(IBM) See abstract, column 3 lines 5-9, column 5 line 53-column 6 line 13	-

X	Document indicating lack of novelty or inventive step	A	Document indicating technological background and/or state of the art.
Y	Document indicating lack of inventive step if combined with one or more other documents of same category.	P	Document published on or after the declared priority date but before the filing date of this invention.
&	Member of the same patent family	E	Patent document published on or after, but with priority date earlier than, the filing date of this application.